ニューラルネットワークを用いた顔表情認識

東北大医保健学科紀要　13(1)：23～32，2004

ニューラルネットワークを用いた顔表情認識

小谷中洋介[1]，本間経康[2]，酒井正夫[3]，阿部健一[1]

[1]東北大学大学院工学研究科　電気・通信工学専攻
[2]東北大学医学部保健学科　放射線技術科学専攻
[3]東北大学大学院情報科学研究科　情報基礎科学専攻

Recognition of Facial Expressions Using Neural Networks

Yosuke KOYANAKA[1], Noriyasu HOMMA[2], Masao SAKAI[3], and Kenichi ABE[1]

[1]*Department of Electrical and Communication Engineering, Graduate School of Engineering, Tohoku University*
[2]*Department of Radiological Technology, School of Health Sciences, Faculty of Medicine, Tohoku University*
[3]*Department of Computer and Mathematical Sciences, Graduate School of Information Sciences, Tohoku University*

Key words: Facial Expressions, Neural Networks, Back-propagation, 2-Dimension Discrete Cosine Transformation

In this paper, we develop a novel recognition system of human face expressions by using neural networks. An essential core of the recognition system is a new categorization method of the facial image data for the neural network learning. The categorization is carried out not based on self-assessment of the person who shows the facial expressions, but on more objective judgements by the 3rd persons. Simulation results show that the proposed system can recognize the facial expressions more easily and accurately compared with conventional method.

1. はじめに

我々が，他者との親密なコミュニケーションをとったり，無用な争いを避けたりすることができるのは，顔に表出されたいろいろな表情を認識し，それを手がかりとして適応的に行動できることに負う部分が大きい[1]。このような顔表情認識の能力をコンピュータシステムに持たせることが可能となれば，ユーザ（人間）とコンピュータシステムとの円滑なコミュニケーションの実現が期待できる。つまり，コンピュータシステムにおける顔表情認識は次世代のヒューマンインターフェイスの構築において重要な研究テーマの一つである。

コンピュータシステムによる顔表情認識に関する研究は古くから行われているが，そのほとんどでは，被験者の無表情と何らかの表情における画像情報の差分（変化量）を処理・解析することで表情を識別する手法が一般的となっている[2,3]。

表情の変化量を使用するために必要な顔表情の定量的記述においては，Ekmanらによって開発された顔面動作記述法（Facial Action Coding System: FACS）が有名である[2]。このシステムは，顔の筋肉の動きを44種の標準動作単位（Action Unit: AU）に分割し，その組合せで任意の顔表情の定量的記述を可能とする。しかし，実

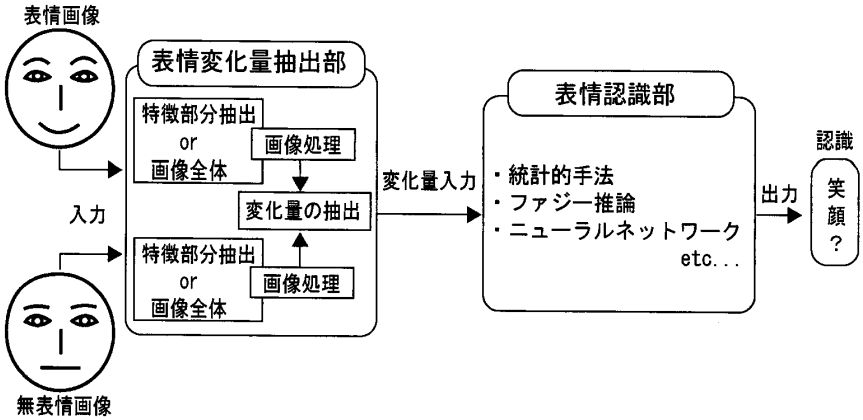

図1.　コンピュータシステムにおける一般的な顔表情認識

際の顔画像認識にFACSを用いる場合には顔の筋肉の動き検出に3次元計測が必要となり，その処理が大変複雑になってしまう。そのため，44種あるAUから表情変化に大きく影響を与える17種のみを用いた研究も報告されているが，撮影条件の異なる顔画像からAUの変化を抽出することは極めて困難であり実用的ではないと考えられる。一方，空間周波数を用いて顔画像の表情変化を抽出する試みも報告されている。主なものは，顔画像の一部または全体に2次元離散余弦変換（2-Dimention Discrete Cosine Transformation：2-D DCT）を施し，その差分を変化量とすることで顔表情認識を行う方法である[3]。その中で肖らは，顔全体の2-D DCTの変化量とニューラルネットワークを用い，顔の特徴部分の抽出を必要としない簡単な表情認識法を提案し，優れた精度で認識が可能であることを示した[4]。

しかし，認識に用いるデータの前処理が多くの計算を要する問題点や，学習に用いている顔表情画像は被撮影者の自己申告により主観的に分類されている。感情情報の認識は，そのような主観的判断によらず，第3者の客観的判断で行うことに妥当性があるという考えがある[5]。

本論文では，計算量を低減しかつデータの取得（撮影）条件に影響を受けにくい新たな前処理法を提案するとともに，学習に用いる顔表情画像をアンケートにより客観的判断で分類を行い，それにより認識率を向上させる新たな認識システムを構築する。簡単なシュミレーション結果より提案システムの有効性を示す。

2.　ニューラルネットワークを用いた顔表情認識法

本論文では，顔表情変化量の抽出において，顔画像全体に2次元離散余弦変換（2-Dimention Discrete Consine Transformation：2-D DCT）を施すことにより空間周波数情報に変換し，その低周波領域のエネルギー変化量を用いる。また，顔表情認識においては，その変化量を入力としてニューラルネットワークを用い，対応する表情の認識を行う手法[4]を用いる。以下に，その詳細を述べる。

2.1.　表情変化量抽出

2-D DCTは，数多く提案されている周波数領域への変換手法の一つである。また，2-D DCTは直交変換の一つで，画像圧縮分野においても広く用いられている。ディジタル画像のピクセルデータは膨大で，その中から表情特徴を抽出することはかなり大変である。2-D DCTを用いることでデータを圧縮可能であり，表情特徴の抽出を容易にする効果が期待できる。2-D DCTの変換式は次式で与えられる。

$$F(k_1,k_2)=\frac{4C(k_1)C(k_2)}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}f(i,j)\times\cos\frac{(2i+1)k_1\pi}{2N}\cos\frac{(2j+1)k_2\pi}{2N}$$

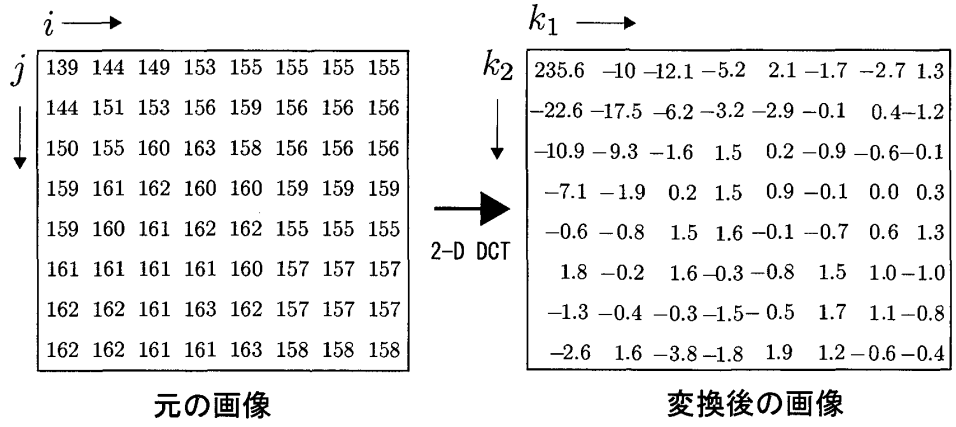

図 2. 画像の数値データの 2-D DCT（128×128 ピクセル中，8×8 ピクセル部分）

$$k_1, k_2 = 0, 1, 2, \ldots, \quad i, j = 0, 1, 2, \ldots, N-1$$

$$C(k) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{if } k = 0 \\ 1, & \text{if } k \neq 0 \end{cases}$$

ここで，$f(i,j)$ は空間データ，(i,j) 座標の値である。式(1)から分かるように，2D-DCT では，余弦成分を抽出するような変換を行う。

一般的な画像の性質として，低周波数成分のエネルギーが大きく，高周波数成分のエネルギーは少ないという特徴がある。そのため周波数領域で表している変換後の画像は，相対的に低周波領域に絶対値の大きな値が集中する（図 2 の変換後の画像の左上部分に相当）。

次に，2-D DCT を施した顔画像から認識部となるニューラルネットワークへの入力データを作成する。

まずはじめに，周波数領域において表情の変化量を抽出するための相対的な尺度を与える。ある人物の無表情顔画像と表情顔画像の 2-D DCT 変換画像をそれぞれ $F_n(k_1, k_2)$, $F_e(k_1, k_2)$ とする。そして，表情変化量 $\Delta F_e(k_1, k_2)$ を次式のように定義する。

$$\Delta F_e(k_1, k_2) \triangleq F_e(k_1, k_2) - F_n(k_1, k_2) \qquad (2)$$

ここで，表情変化量の低周波領域のみの $k_1, k_2 = 0, 1, 2, \ldots, I$ を認識部で使用するデータの対象範囲とし，これを更に $B \times B$ $(B < I)$ の要素からなる小ブロック（図 3）に分けて，その $B \times B$ 行列の平均値をデータの値とする。またこれにより，認識部で使用するデータ数は $(I/B)(I/B+1)/2$ となる。これより，認識部への入力データ $\boldsymbol{X}$ を，図 3 のように次式で定義する。

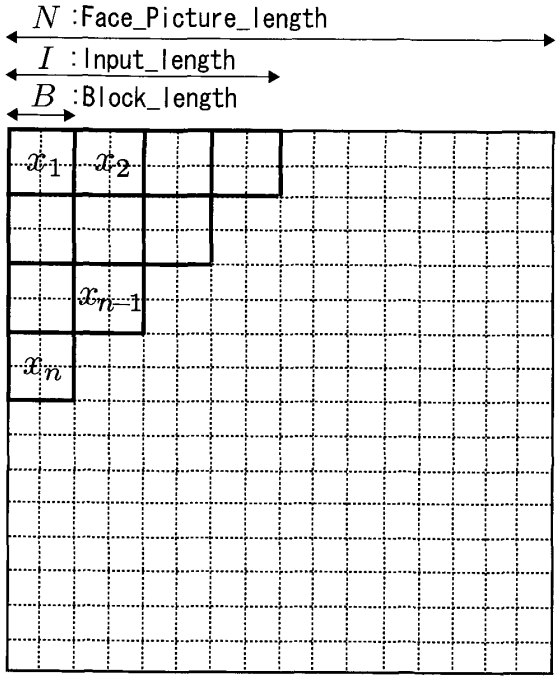

図 3. 表情変化量のブロック分割

$$\boldsymbol{X} = [x_1\ x_2 \ldots x_n],\ n = \frac{\frac{I}{B} \cdot \left(\frac{I}{B}+1\right)}{2} \qquad (3)$$

なお本論文では，$I = 16$，$B = 1$ と設定し，デー

タ数は136個とした。

2.2. 表情認識

表情変化量抽出部で得られた入力データを認識部となるニューラルネットワークへ入力する。2-D DCTによる表情変化量から適当な表情への認識は数学的に処理することは困難であるが，ニューラルネットワークのもつ自己学習能力を利用することにより，比較的容易に処理することができる。表情認識におけるニューラルネットワークのメカニズムは，パターン認識そのものであり，本論文ではパターン認識に適した図4のような3層構造のフィードフォワードネットワーク(Feed-forward Neural Networks: FNN)を用いる。学習アルゴリズムには，モーメンタムを有する誤差逆伝播(Back-propagation: BP)法を使用する。モーメンタムを用いたBPは，浅い極小値に入り込むことを避け，より深い極小値をとらえる可能性が高いことが知られている。ネットワークのシナプス重みと閾値の行列 W は，

$$\boldsymbol{W}(t+1)=m_c\boldsymbol{W}(t)+(1-m_c)\beta\Delta\boldsymbol{W}(t) \quad (4)$$

のように更新される。ここで，ΔW はシナプス重みと閾値の変化量であり，β は学習係数，m_c はモーメンタム定数である。

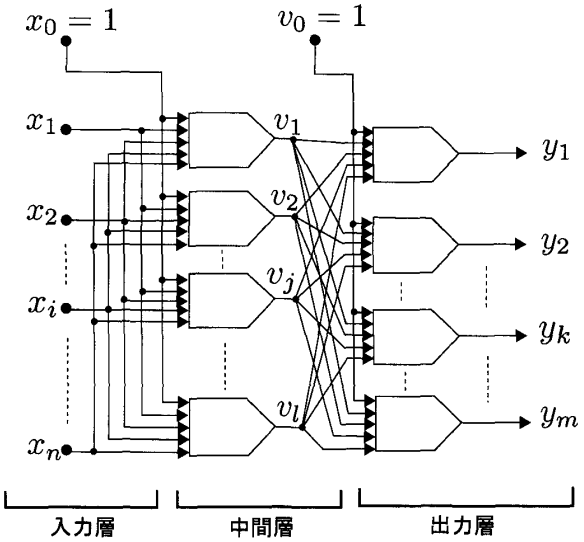

図4. ニューラルネットワークの構造（FNN）

2.3. 教師信号

今回の手法において認識する対象表情は，「笑顔」・「怒り」・「驚き」・「悲しみ」の4表情である。この4表情に対するニューラルネットワークの目標とする出力（教師信号）を，以下のような2値の4次元ベクトルとして定義する。

$$\begin{aligned}
&\text{笑顔:} \quad T=[\ 1 \quad -1 \quad -1 \quad -1\ ]\\
&\text{怒り:} \quad T=[-1 \quad 1 \quad -1 \quad -1\ ]\\
&\text{驚き:} \quad T=[-1 \quad -1 \quad 1 \quad -1\ ]\\
&\text{悲しみ:} \quad T=[-1 \quad -1 \quad -1 \quad 1\ ]
\end{aligned} \quad (5)$$

本論文では，学習用の表情データに対して，BP法を用いてニューラルネットワークを学習させた後，未学習の表情データに対して表情認識を行う。ニューラルネットワークによる認識表情の決定は，抽出した表情変化量をネットワークに入力し，その出力値の中で最も大きい数値を示す出力を該当の顔表情とする。ただし，それが2番目の最大値との差が小さくて，次の条件（ネットワークの最大出力と第2位の出力との差が最大出力の1/3以上）を満たさない場合にはその顔画像を認識不可能（拒否）と判断するものとする。

$$Y_M-Y_S\geq\frac{Y_M}{3}$$

Y_M: 最大出力

Y_S: 第2位の出力 (6)

3. 提　案　法

3.1. 正規化法

本論文で用いた顔表情データベースは，20代の男女30人による5つの顔表情（無表情・笑顔・怒り・悲しみ）による合計150枚の顔画像（白黒，128×128ピクセル，256階調）から構成される。

また，照明や画像中の顔の位置などの撮影条件を完全に一致させることは困難であるため，正確な差分を計算することはできず，そのままでは認識精度が低下してしまう問題がある。そこで，サンプル同士のズレを補正するために，顔画像の正規化を行うのが一般的である。正規化における重要な処理に参照点（両目・口）の抽出がある。本論文では，正規化の処理において，一般的に用い

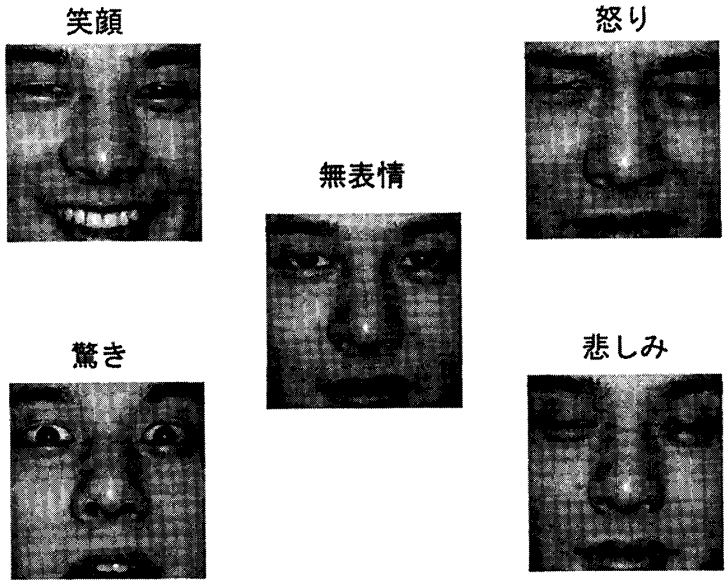

図5. 顔表情画像の一例

られる赤松らの手法[6]を参考とし，より簡単でかつ背景や照明などの撮影条件にあまり影響されない方法を提案する。

赤松らの手法では，RGBカラー画像から3種の変換画像（色相を表す画像，カラーテレビ信号における色差信号，I信号画像とQ信号画像）を作成し，さらに多種の画像処理を行うことで両目・口の抽出を行っている（図6）。この参照点を用いてアフィン変換を行うことにより，白黒の正規化画像を作成する。しかし，この手法では，画像の背景や照明具合などに強く影響するためうまく抽出できない場合もある。更に，多種の画像変換・処理を行っているために計算が複雑となっている。

提案法では，一般に人物肌色の彩度値は比較的高いことに着目し，そのような変換を施したQc画像において値の大きい領域を抽出すれば，顔領域のみを抽出することができる[7]という性質を利用している。図7が，本論文における両目・口の抽出法の流れ図であり，以下にその詳細について示す。

(i) まず，256階調のRGBカラー画像である元画像をQ_c画像に変換する。このQ_c画像は画像の彩度（Quasi-chroma）を表す。変換式は次式で与えられる。

$$Q_c=\sqrt{\left(\frac{2R-G-B}{2}\right)^2+\left(\frac{\sqrt{3}(G)-B}{2}\right)^2} \tag{7}$$

(ii) 抽出した顔領域の画像を，カラーテレビ信号の色差信号の一つであるQ信号画像に変換する。Q信号画像は緑からマゼンダ（赤と青の中間色）への信号となり，マゼンダに近くなるほど値が大きくなる。変換式は次式である。

$$Q=0.12R-0.52G+0.31B \tag{8}$$

一般的に人間の唇の色は，赤とマゼンダの間であるため，このQ画像の値の大きい領域を抽出すれば，口の位置を得ることができる。

(iii) 次に，両目の抽出を行う。今度は，Qc画像のエッジ検出を行う。エッジ検出は，微分による輪郭抽出法で，色の急激な変化の部分を検出することができる。Qc画像を見ると分かるが，目の部分は顔の中で比較的特異な色となり，急激な色の変化が起きている。これにより，両目の領域を抽出することができる。しかし，眉や前髪が垂れている場合も，その領域では比較的大きな

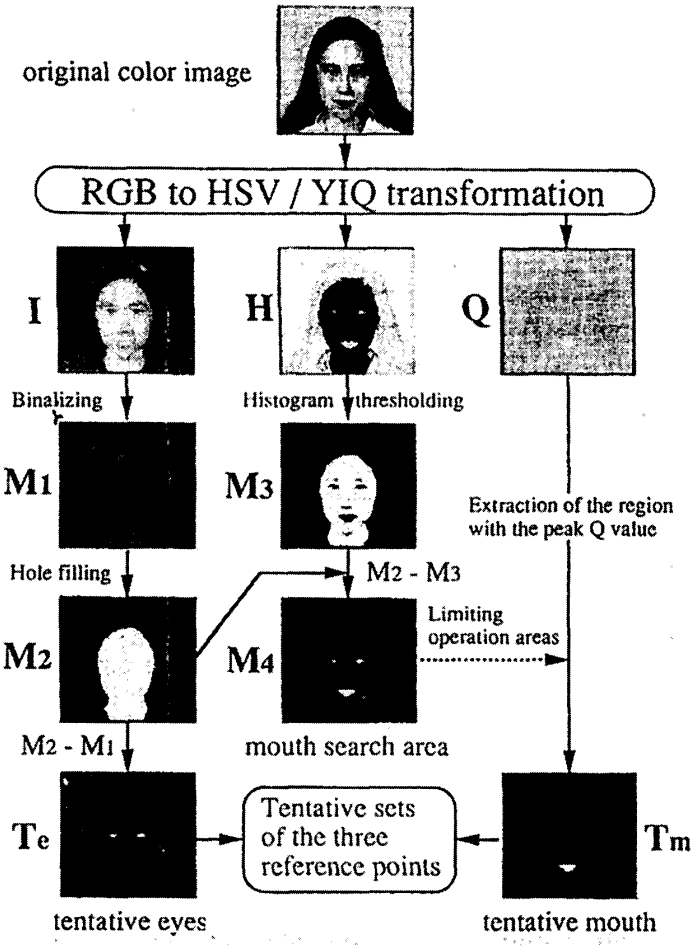

図 6.　赤松らの手法[6]による抽出法

色の変化となる。そのため，両目の抽出は困難となる。ここで，人間の両目と口の位置関係は一般的に大差がないことから，口の位置よりエッジ検出画像から両目の探索領域を指定する。この補助的情報を用いることで，両目の抽出が行える。

上述の方法を用いて，両目・口の位置を抽出し，それらを参照点として正規化を行う。本論文では，両目の中心点と鼻との距離を d とし，口の中心から画像の左右の隅への長さを $c_1=c_2=0.8d$，目の中心から画像の上隅への長さを $c_3=0.4d$，下隅への長さを $c_4=0.8d$ となるような正規化画像（白黒，128×128 ピクセル，256 階調）を使用した（図 8）。

3.2.　顔表情データベースの再編

一般的に用いられている顔表情画像データは被撮影者の自己申告，つまり主観的カテゴリを用いて分類している。しかし，被撮影者の自然な感情に基づく顔表情を取得（撮影）しているとは限らず，裏付けとなる感情なしに，いわば人工的かつ強制的に創出した顔表情を撮影せざるを得ない場合も多い。人工的に創出された表情は，他人からはその表情どおりには見えず，他の表情に見えたり，またはどの表情にも見られない可能性がある。

本論文では，この問題を客観的に確認するために次のような調査を行った。調査法は，「被撮影者の自己申告で分類された無表情画像と各表情画

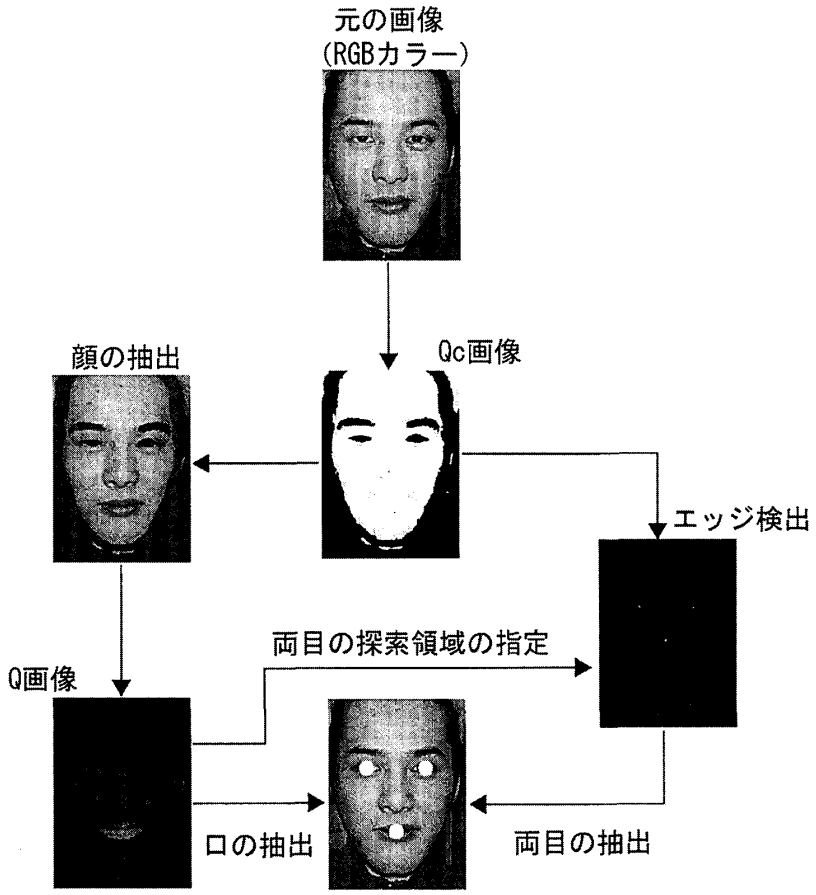

図7. 提案抽出法

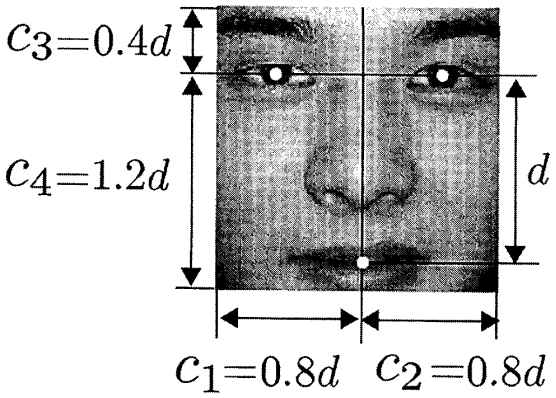

図8. 正規化画像

像を被験者に見比べてもらい，その表情がどの表情に見えるか?」というアンケート方式である。今回は，顔表情データベースより60表情(15人4表情）に対して調査を行った。アンケートの回答者は20代の男女20人で，回答総数中7割以上の支持を得た表情をその顔画像の客観的表情とし，どの表情にも7割以上の回答を得られなかった場合は判断不能として「拒否された表情」とした。集計結果を表1に示す。

縦の表情が被撮影者の自己申告による表情で，横の表情がアンケートで7割以上の回答をを得た表情である。この結果から分かるように，「笑顔」に関しては，被撮影者による主観的表情とアンケート結果による客観的表情が一致している。しかし，他の3表情に関しては，半分にも満たない数の表情しか主観的表情と客観的表情は一致しなかった。さらに，「驚き」に関しては，客観的表情において「笑顔」として見られたのもあった。このアンケートの結果より，被撮影者の判断による主観的表情というものは必ずしも他人に認識される客観的表情とは一致しないということが確かめられた。このことは，学習に用いられる教師信号

表 1. アンケートの結果

アンケート結果による表情 ＼ 自己申告による表情	笑　顔	怒　り	驚　き	悲しみ	拒　否
笑　顔	15	0	0	0	0
怒　り	0	6	0	0	9
驚　き	3	0	7	0	5
悲しみ	0	0	0	5	10

表 2. アンケートによる表情の客観的分類

	被撮影者による主観的分類	アンケートによる客観的分類
笑　顔	30	32
怒　り	30	17
驚　き	30	19
悲しみ	30	17

に誤りがあることに相当するため，たとえ認識対象である各種表情パターン群自体が，特徴要素空間において識別（分類）可能である場合でも，ニューラルネットはそれらのねじれた関係を学習してしまう可能性があることを意味している。このねじれより，正常な認識が行えない可能性がある。そこで本論文では，顔表情認識に使用する顔表情データベースをアンケートによる客観的分類で再編を行うことにする。アンケートの回答者は20代男女20人で，回答総数中7割以上の支持を得た表情をその顔表情画像の客観的表情とし，どの表情にも7割以上の回答を得られなかった場合は拒否された表情とした。表2に，120表情（30人4表情）をアンケートによる客観的分類で判別した結果を示す。

すべての表情で，データ数を均一にするため，以下のシミュレーションで用いるデータ数は4表情（×17人分）とする。

3.3. 再編データベースによる顔表情認識

再編した顔表情データベース（客観的分類）を用いて行った表情認識のシミュレーション結果を示す。今回は17人分の顔表情画像中，11人分の

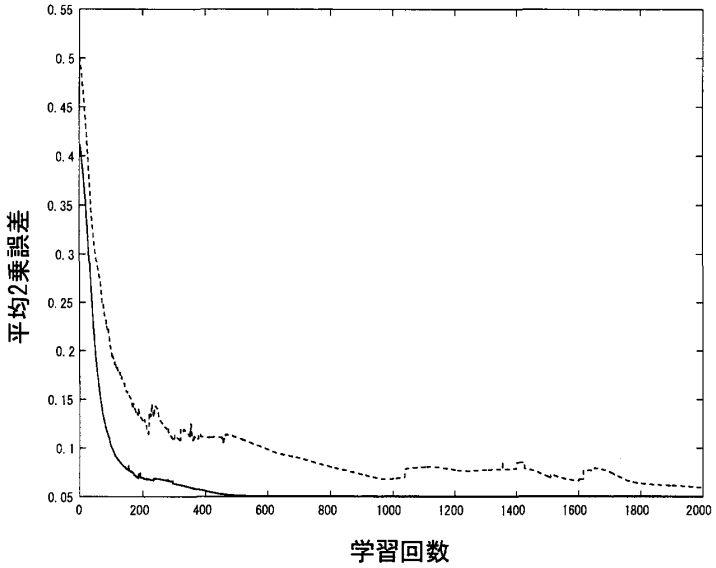

図 9. 再編データによる学習曲線

表 3. 被撮影者の主観的分類による表情認識結果

	笑　顔	怒　り	驚　き	悲しみ	拒　否	認識率
笑　顔	2	1	0	0	3	33%
怒　り	0	3	0	0	3	50%
驚　き	1	1	0	0	4	0%
悲しみ	1	1	0	0	4	0%

表 4. アンケートの客観的分類による表情認識結果

	笑　顔	怒　り	驚　き	悲しみ	拒　否	認識率
笑　顔	4	0	0	0	2	67%
怒　り	0	3	1	0	2	50%
驚　き	0	0	3	1	2	50%
悲しみ	0	0	1	1	3	17%

顔表情画像を学習に使用し，残り 6 人分の顔表情画像を学習後の表情認識に用いた。また，ニューラルネットワークにおけるパラメータは，学習定数 $\beta=0.01$, モーメンタム定数 $m_c=0.95$, 中間層の素子の数は 30 個と設定した。また，中間層と出力層のシナプス重み・閾値の初期値はランダムで決定する。今回は比較のために，被撮影者による主観的分類を用いたデータベース（17 人 4 表情）を用いた場合の認識も行った。まず，学習曲線を図 9 に示す。破線が被撮影者による主観的分類によるデーターベースを用いた場合の学習曲線，実線がアンケートによる客観的分類による学習曲線である。

被撮影者による主観的表分類のデータベースにおける学習の平均 2 乗誤差は 0.06 であるが，対してアンケートによる客観的分類の方は平均 2 乗誤差が 0.03×10^{-3} となっている。また，誤差の収束速度も後者の方が早い。このことから，アンケートによる客観的分類を用いた方が各表情に共通性が生じ，学習が容易になることが分かる。

次に，学習後の表情認識の結果を表 3, 4 に示す。これらの結果から分かるように，認識対象を被撮影者の主観的分類からアンケートの客観的表情にすることにより，表情認識の認識率に向上が見られる。このことは，学習に用いるデータ（顔表情）画像をカテゴリ分類する際に，被撮影者の主観のみで判断せず，なんらかの客観的評価が必要であることを示唆している。被撮影者は，その顔表情を創出する感情的刺激がない状態で，人工的に，いわば無理にその顔表情を創出することを強いられていることから考えても，妥当な考察結果であると考えられる。逆に実際に，あるいは VR 技術を用いて感情的刺激を被撮影者に与えることにより実際の顔表情に即した認識システムを構築することが可能である。

4. むすび

本論文では，顔表情認識システムにおいて計算量を低減しつつ撮影条件に影響されないデータの正規化法を提示した。また，被認識者の主観による不適切学習を避けることが可能な新しい認識システムを提案した。しかし，今回使用した顔画像の客観的な表情の分類は，20 人程度によるアンケートで行ったものであり，一般的な客観的分類とは言い難い。したがって，判断する人数を増やすなどより複数人の判断により客観的に分類したサンプルを用いて表情認識を行う必要があると考えられる。

従来方との比較により提案手法の有効性は示されたが，今回得られた認識率は満足のいくレベルではない。しかし，同じ認識法を用いても単純に顔表情データ画像数を多くすることで，認識率が向上することを確認している。サンプル画像数を増やし，より高精度な認識システムを提案することは今後の課題である。

文　献

1) 千葉浩彦 他：顔と心―顔の心理学入門―，サイエンス社，1993
2) P. Ekman and W. Friesen: Facial Action Coding System, Consulting Psychologists Press, San Francisco, U.S.A, 1977
3) K. Ebihara, J. Ohya and F. Kishino: A study of real time facial expression detection for virtual space teleconferencing, IEEE International Workshop on Robot and Human Communication, pp. 247-252, 1995
4) 肖　業貴，N.P. チャンドラシリ，田所嘉昭，尾田正臣：2-D DCT とニューラルネットワークを用いた顔画像の表情認識，電子情報通信学会論文誌，Vol. J81-A, No. 7, pp. 1077-1086, 1998
5) 佐藤　他：音声に含まれる感情情報の認識に関する研究，ヒューマン情報処理研究会資料，2001
6) S. Akamatsu, T. Sasaki, H. Fuamachi and Y. Suenaga: Automatic extraction of target images for face identification using the subspace classification method, IEICE Trans. Inf & Syst, vol. E76-D, pp. 1190-1198, 1993
7) 平山泰崇，中村　納：色情報と等濃線分布に基づいた顔画像による人物識別方式，テレビジョン学会技術報告，Vol. 20, No. 41, pp. 7-12, MIP96-54, 1996